Alex Voorhoeve

Heuristics and biases in a purported counterexample to the acyclicity of "better than"

Working paper

Heuristics and Biases

in a Purported Counterexample to the Acyclicity of "Better Than".[1]

Alex Voorhoeve

Philosophy, LSE.

a.e.voorhoeve [at] lse.ac.uk

Draft, June 2007. Please do not quote or circulate without permission.

Abstract

Stuart Rachels and Larry Temkin have offered a purported counterexample to the acyclicity of the relationship "all things considered better than". This example invokes our intuitive preferences over pairs of alternatives involving a single person's painful experiences of varying intensity and duration. These preferences, Rachels and Temkin claim, are confidently held, entirely reasonable, and cyclical. They conclude that we should drop acyclicity as a requirement of rationality.

I argue that, together with the findings of recent research on the way people evaluate episodes of pain, the use of a heuristic known as similarity-based decision-making explains why our intuitive preferences may violate acyclicity in this example. I argue that this explanation should lead us to regard these preferences with suspicion, because it indicates that they may be the result of one or more biases. I conclude that Rachels' and Temkin's example does not provide sufficient grounds for rejecting acyclicity.

Keywords: Acyclicity, rationality, similarity-based decision-making, heuristics and biases, pain evaluation.

Word count (including abstract and endnotes): 6,500.

Heuristics and Biases

in a Purported Counterexample to the Acyclicity of "Better Than".

The acyclicity of "better than" requires that when an alternative a_1 is better than a_2, …, a_{i-1} is better than a_i, …, a_{n-1} is better than a_n, then a_n is not better than a_1, irrespective of the length n of the chain. Acyclicity is a central principle of rational choice, as on a finite set of feasible alternatives, it is necessary for the existence of an alternative which is not worse than some alternative in that set.[2] If, as standard rational choice theory holds, rationality requires not choosing an alternative that is worse than some other feasible alternative, then there is no basis for rational choice when acyclicity is violated.

Recently, Stuart Rachels and Larry Temkin have offered a purported counterexample to the acyclicity of the relationship "all things considered better than".[3] This example invokes our intuitive preferences over pairs of alternatives involving a single person's painful experiences of varying intensity and duration. Rachels and Temkin claim that these preferences are confidently held, entirely reasonable, and cyclical. They conclude that we should not regard acyclicity as a requirement of rationality.

In this paper, I defend the acyclicity of "all things considered better than" against this example. No matter how intuitively and theoretically attractive a principle like acyclicity is, our attachment to it may be undermined by its lack of fit with considered and confidently held case judgments. I will argue, however, that the intuitive judgments evoked by Rachels' and Temkin's example should not be confidently held. I review their example in section 1. In section 2, I argue that it elicits the use of a heuristic known as "similarity-based decision-making". I also argue that, together with the findings of recent research on the way people evaluate episodes of pain, the use of this heuristic explains why our intuitive preferences in this example may violate acyclicity.[4]

In section 3, I argue that this explanation should lead us to regard these preferences with suspicion, because it indicates that they may be the result of incorrectly weighing the intensity of pain or the duration of pain in at least one of the comparisons that Rachels and Temkin ask us to make.

1.

Rachels and Temkin propose several counterexamples to acyclicity involving pains of varying intensity and duration. These examples are sufficiently similar that we can safely focus on only one of them. Imagine the prospect of living for a further substantial, fixed number of years T in good health and without any significant pain, except for the fact that you will have to endure a certain episode of pain E_i which will begin tomorrow. Assume that however bad this episode is, it is never so bad as to render the period T as a whole not worth living. To begin with, imagine that this episode is a significant, though compared to T relatively short period of excruciating torture. Call this combination of intensity of pain and the time it must be endured E_0. Now imagine enduring slightly less intense torture for much longer than the duration of the pain in E_0. Call this combination of intensity of pain and the time it must be endured E_1. Though the prospect of enduring E_0 is awful, it will intuitively seem better, Rachels and Temkin claim, to endure E_0 than to endure E_1. Now consider E_2, which involves suffering a pain slightly less intense than the pain in E_1 for much longer than the duration of the pain in E_1. Again, Rachels and Temkin argue, E_1 seems intuitively better than E_2. Now iterate this kind of reasoning and by so doing construct a sequence of two-dimensional alternatives $E_0, E_1, \ldots, E_{MILD}$ in which the final member E_{MILD} involves a mild pain (such as the pain of a mild headache or a hangnail) for a long period of time which is less or equal to the time T you have left to live. It is possible, they

claim, to construct this sequence so that in the pairwise comparison of adjacent members, the slight alleviation of the intensity of pain always appears to be outweighed by the marked increase in its duration, so that each member in this sequence will be preferred to its successor. Now, Rachels and Temkin argue, if the period of excruciating torture is long enough, it is also intuitively preferable to endure a mild pain for a very long time than to face this significant period of torture, so that E_{MILD} is intuitively better than E_0. In sum, in pairwise comparison, each member of this sequence is preferred to its successor and the final member of this sequence is preferred to the initial member, violating acyclicity.[5] Rachels and Temkin believe that even after reflection, it is perfectly reasonable to confidently regard E_0 as better than E_1, E_1 as better than E_2, etc. and E_{MILD} as better than E_0. They conclude that "all things considered better than" is not an acyclic relation.

Rachels and Temkin offer the following explanation for the purported failure of acyclicity in this case.[6] Which features of a given alternative are relevant, and/or how significant these features are, may depend on which alternative it is being pairwise compared with. Consequently, though an alternative E_0 may be better than E_1 and E_1 may be better than E_2 in terms of the features that are relevant to these comparisons and the significance that these features have when making these comparisons, E_0 need not be better than E_2 in terms of the features of these alternatives that are relevant to that comparison or the significance that these features have in that comparison. Rachels and Temkin argue that their example involves precisely this kind of shift in the apparent relevance and/or significance of the features of the alternatives. The duration of a painful experience is of great importance, they argue, when we compare two experiences which differ only slightly in intensity. However, when we compare two experiences which differ greatly in intensity, the duration of the experience is not always so significant. Temkin puts this point as follows:

> "In comparing pains that merely differ in degree, duration clearly plays a significant role. That is why we think that a shorter intense pain might clearly be better than a much longer less intense pain. But in comparing pains that differ in kind, duration plays a very different role. In comparison with torture of sufficient duration, a hangnail's duration basically does not matter. So, a factor that is clearly relevant and significant in comparing some outcomes is not relevant—or at least has very different significance—in comparing different outcomes. Thus, [acyclicity] fails for reasons that are clear, straightforward, and, I think, perfectly appropriate."[7]

One natural response to this argument is to question the reliability of the intuitive judgments involved in the following way. The example involves very imprecisely described intensities of pain—except for the first few and last few members of the sequence, the intensities of pain are not described in absolute terms, but only in terms like "slightly less intense than the pain experienced in the preceding alternative". Moreover, most people will have little to no experience of the intensities of pain under discussion. The durations involved are also imprecisely indicated. This lack of clarity about the nature of the alternatives helps obscure that the construction of a sequence of the kind Rachels and Temkin have in mind is less straightforward than it may first appear to be. For if one makes the differences in intensity of pain between adjacent members of this sequence very small in order to render the preference for the earlier member compelling, then it will take a great many steps to arrive at a mild pain, and the durations one would need to invoke in the final members of the sequence will be longer than we can properly appreciate. On the other hand, if one makes the differences in intensity of pain sufficiently large to get from excruciating torture to mild pain in a number of steps that is small enough to ensure that the durations involved are not excessive, then the case for always preferring the earlier member of any two adjacent members of the sequence may not be

compelling. Since we have insufficient evidence to believe that we have confidently held preferences in this case that violate acyclicity, this response concludes, we have insufficient grounds for abandoning acyclicity.[8]

I am sympathetic to this response. However, it leaves unexplained several striking features of some people's experience when confronted with this example. First, it leaves unexplained why we may feel intuitively drawn to express preferences with a high degree of confidence over alternatives which are so vaguely described and the features of which are so unfamiliar to us. This is striking, because it might seem more natural for us simply to state that we have no idea how we would evaluate such alternatives. Second, it leaves unexplained why we may feel drawn to believe that it will be possible to construct a sequence over which these intuitive preferences will violate acyclicity without making use of inordinately long periods of time. (For example, I believe one could easily imagine having cyclical preferences of the hypothesized kind over a sequence like the one outlined in Table 1—where the final episode lasts for around 39 years.) Third, it does not account for Rachels' and Temkin's explanation of this violation of acyclicity in terms of the apparently changing relevance and/or significance of the two features of the alternatives in question.

Table 1. An imaginary example of a sequence of the Rachels-Temkin kind.

Episodes	Attributes	
	Intensity of pain*	Duration in weeks
E_0	20	1
E_1	18	2
E_2	16	4
E_3	14.5	7
…	…	…
E_{21}	2.0	650
E_{22}	1.5	1100
E_{MILD}	1.0	2050

*This is assumed to be a cardinal scale.

I believe these facts can be explained by turning to recent work on two topics in intuitive decision-making: similarity-based decision-making and the evaluation of painful experiences. I will discuss each in turn.

2.

Ariel Rubinstein's influential characterization of similarity-based decision-making runs as follows.[9] When deciding between multi-dimensional alternatives, say bundles of pain-intensity and the time it must be endured (p_i, t_i) and (p_j, t_j), a decision-maker goes through the following three-stage procedure:

> *Stage 1*: The decision-maker looks for dominance. If $p_i < p_j$ and $t_i < t_j$, then bundle (p_i, t_i) is preferred to bundle (p_j, t_j).
>
> *Stage 2:* The decision-maker looks for similarities between p_i and p_j and between t_i and t_j. If she finds similarity in one dimension only, she determines her preference

between the two pairs using only the dimension in which there is no similarity. For example, if p_i is similar to p_j while t_i is not similar to t_j, and $t_i<t_j$, then bundle (p_i, t_i) is preferred to bundle (p_j, t_j).

Stage 3: The choice is made using an unspecified different criterion.

Recent tests have yielded significant evidence supporting the hypothesis that people make decisions in this way or in closely related ways in a variety of choice situations including gambles (with prizes and the probability of winning them as dimensions of the alternatives), the choice of applicants (where the dimensions were taken to be "intellectual ability", "emotional stability", and "social facility"), inter-temporal trade-offs (where the dimensions were time and money), and the choice of jobs (where the dimensions were commuting time and the wage level).[10]

Subjects are hypothesized to use similarity-based decision-making because it simplifies decision-making in various ways. First, by using information on similarities and differences, it draws on easily accessible knowledge. Similarity appears to be among the features of objects that are routinely and automatically registered by the perceptual system; we also appear to be better attuned to the evaluation of differences than to the evaluation of absolute magnitudes.[11] Second, by placing intra-dimensional evaluation before the possible use of inter-dimensional evaluation, the procedure makes use of the fact that intra-dimensional evaluation is simpler, because it involves comparisons between features of alternatives that are expressed in the same units.[12] Finally, it ensures that it will be immediately apparent when an alternative is slightly better than another along all relevant dimensions, a fact which might be obscured if the overall goodness of each alternative was first evaluated independently.[13]

Though similarity-based decision-making will not necessarily yield preferences that violate principles of rational choice, the use of similarity-based decision-making can explain a wide range of violations of axioms of orthodox decision theory, including violations of acyclicity.[14] It is, for example, a common explanation for why experimenters manage to get subjects to reveal cyclic

preferences when they present subjects with a sequence of pairwise choices between multi-dimensional alternatives with the following properties: (i) each alternative in the sequence is always slightly better than its predecessor along the first dimension and markedly worse along the other dimension(s); (ii) when the difference between alternatives along the first dimension is substantial, this dimension becomes especially significant. In a sequence of this kind, when someone using similarity-based decision-making compares adjacent alternatives, the slight improvement along the first dimension always appears to be outweighed by the marked worsening along the other dimension(s), so that each alternative is preferred to its successor in the sequence. However, the same person, when comparing the initial and final alternatives in the sequence will find that the sequence of slight improvements along the first dimension adds up to a substantial improvement along that dimension. Given the hypothesized importance of the first dimension when alternatives differ markedly along it, this substantial difference along the first dimension intuitively outweighs the large cumulative worsening along the other dimension(s), so that the final alternative is preferred to the initial alternative.

An experiment of Amos Tversky's can serve as an example. Tversky asked students at Harvard University to make pairwise choices between potential applicants who were characterised by their percentile ranks along three dimensions, "intellectual ability" (I), "emotional stability" (E), and "social facility" (S).[15] During the experiment, subjects were asked to make pairwise choices between the candidate profiles presented in Table 2, among others. The results indicated a significant proportion of subjects violated acyclicity, favouring earlier over later profiles when choosing between adjacent profiles in the sequence *a*, *b*, *c*, *d*, *e*—because the difference in intellectual ability between adjacent profiles appeared slight, and the difference in the other dimensions substantial—while also favouring *e* over *a*—because when the difference in intellectual ability was substantial, this was taken to be decisive, intelligence being considered the most important characteristic for entry to the university.

Table 2. Candidate profiles for one of Tversky's experiments.

Candidate profiles	Dimensions		
	I	E	S
a	69	84	75
b	72	78	65
c	75	72	55
d	78	66	45
e	81	60	35

It is easy to see why Rachels' and Temkin's example may elicit the use of similarity-based decision-making in the pairwise comparison of adjacent alternatives in their sequence E_0, E_1, E_2, …, E_{MILD}. First, given the characterisation of the intensity of pain in each member of the sequence as "just slightly less than" the intensity of pain in the preceding member of the sequence, adjacent alternatives may well be regarded as similar along the intensity of pain dimension, while the marked increase in duration as we move through the sequence ensures that they will be experienced as dissimilar along the duration of pain dimension. Second, the information provided invites the use of similarity-based decision-making, since with the exception of the first and final members of the sequence, we are offered only rough information about the differences in the intensity and duration of pain between adjacent members, and are therefore not informed about the absolute intensity and duration of pain involved in the intermediate alternatives. (Though even if we were offered a more precise description of the alternatives, we would still be disposed to rely on perceptions of similarity and on our evaluations of differences, which are easily accessible, rather than on our appreciation of absolute magnitudes of intensity of pain and the length of time it must be endured, about which we feel less certain.) Third, we might be drawn to similarity-based decision-making because it enables us to avoid the

difficult task of specifying precisely how we should make trade-offs between the intensity of pain and its duration.

Similarity-based decision-making would, of course, lead one to always prefer the earlier of any two adjacent alternatives in Rachels' and Temkin's sequence, which is in line with Rachels' and Temkin's claims about our intuitive preferences. It would not, however, lead to any particular choice between E_0 and E_{MILD}. What does psychological theory have to say about this choice? Recent studies indicate that the relative weight given to the intensity of pain and its duration in the intuitive evaluation of painful episodes depends markedly on the attention directed towards each of these two attributes, and on their evaluability, with the weight given to an attribute increasing in both the amount of attention directed to it and its evaluability.[16] They also indicate that when subjects' attention is not specifically directed to duration, and its contribution to the badness of a painful episode is not made especially easily evaluable, it is given very little weight. In an experiment carried out by Donald Riedelmeier and Daniel Kahneman, for example, patients undergoing a colonoscopy reported the intensity of pain every 60 seconds during the procedure and subsequently provided an evaluation of the total pain suffered during the episode. Each patient underwent one procedure; the length of the procedures varied from 4 to 66 minutes.[17] The task of evaluating the level of pain at regular intervals, coupled with significant changes in the intensity of pain during the procedure, focused patients' attention on the intensity of pain. Moreover, the task of regularly evaluating the intensity of pain rendered the intensity of pain easily evaluable by the end of the episode. By contrast, within the context of the experiment, no subject rated or experienced procedures of different durations. In their global evaluations of each episode, subjects displayed a phenomenon known as "duration neglect": the duration of experiences had little or no independent effect on the way they were evaluated. Instead, subjects appeared to evaluate these episodes by a constructed "representative moment": a collage of the intensity of pain at several singular instants, including the peak and end of the episode.[18] Since the

duration of the episode is not included in this representation, duration was neglected in its overall evaluation.

Duration neglect does not occur, however, in contexts of choice which direct attention towards duration and which render it more easily evaluable. This was illustrated by a series of experiments in which subjects were each exposed to several episodes that contained various unpleasant experiences which differed significantly along three dimensions: duration, intensity, and the direction of intensity over time (increasing, decreasing, or oscillating). The fact that each subject experienced episodes of markedly different duration drew some attention to duration and rendered it more easily evaluable, since it offered subjects points of comparison in terms of duration. In these experiments, subjects appeared to rely on a kind of "anchoring and adjustment" heuristic in evaluating episodes: they took the aforementioned representative moment as a base for their evaluation, and then made significant, but relatively small adjustments to this base to account for the episodes' duration.[19]

It is noteworthy that subjects whose evaluation of painful episodes is heavily determined by their assessment of a representative moment will have preferences that violate normative principles for the evaluation of painful episodes. One such principle is temporal monotonicity, which holds that adding a period of pain to a given painful episode should make it worse. This principle is violated by such subjects, because they will judge a shorter episode of pain as *worse* than a longer episode of pain which contains all the painful experiences of the shorter episode with some additional painful experiences, but which ends on a less unpleasant note.[20] This means that preferences expressed in contexts that elicit duration neglect or the use of the anchoring and adjustment heuristic should be regarded with suspicion.

Duration can be expected to be given greater weight than in the anchoring and adjustment model when subjects are asked to compare experiences that are similar on all dimensions other than duration. For the alternatives' similarity in all other dimensions will make

duration highly salient, and also render differences in duration easily evaluable.[21] (This conclusion is, of course, consistent with the use of similarity-based decision-making in such cases.)

What follows from these findings for our analysis of Rachels' and Temkin's thought experiment? In comparing E_0 and E_{MILD}, our attention will undoubtedly focus on duration, and it will therefore be given significant weight. However, duration will not be as salient or as easily evaluable as it is in the choice between adjacent alternatives in Rachels' and Temkin's sequence. Duration can therefore be expected to receive relatively less weight in the choice between E_0 and E_{MILD} than it does in the choices between adjacent alternatives. This explains why it may be possible to generate a sequence of the required kind: while duration may always have sufficient weight in the comparison of adjacent alternatives to render the earlier alternative preferable, the diminished relative weight of duration in the comparison of E_0 and E_{MILD} means that it may not have sufficient weight to render E_0 preferable to E_{MILD}.

Together, then, similarity-based decision-making and recent research on the evaluation of painful episodes can explain people's responses to Rachels' and Temkin's example. They explain why we may feel intuitively drawn to express preferences with some degree of confidence over alternatives which are so vaguely described and the features of which are so unfamiliar to us: in order to come up with a judgment, the forms of intuitive decision-making that the example elicits do not really make use of the information that is left out, or that we feel uncertain about. Furthermore, they explain the phenomenology of evaluation that Rachels and Temkin describe, in terms of the shifting weight of duration and intensity of pain in our decision-making. This shifting weight is, of course, also the reason why we may find it possible to construct a sequence in which our intuitive preferences violate acyclicity.

3.

Imagine that our preferences over some set of precisely described alternatives, like the options in Table 1, display the pattern that Rachels and Temkin envisage. I believe that if the aforementioned explanation of our preferences is correct, we should regard some of our intuitive preferences with suspicion.

First, a person using similarity-based decision-making while comparing two adjacent alternatives E_i and E_{i+1} does not first consider how bad it would be to endure E_i and then how bad it would be to endure E_{i+1}, and then judge the former less bad than the latter. Instead, alternatives are compared aspect-by-aspect, with the dissimilar aspect always proving decisive. Thus, the decision-maker never explicitly considers how to trade off duration against intensity of pain, or how these two aspects together contribute to the overall badness of an alternative. Finally, the lack of attention directed at intensity of pain in pairwise comparisons may imply that this dimension is being underweighted in some pairwise comparisons.[22] These considerations cast doubt on the validity of our intuitive preferences between adjacent alternatives in Rachels' and Temkin's sequence.

Second, the surprisingly small role of duration in the overall evaluation of painful episodes in some contexts of choice indicates that there is a distinct possibility that duration will be underweighted in the choice between E_0 and E_{MILD}.

These worries about the trustworthiness of some of our intuitive preferences should increase once we realise that our preferences will also violate the principle of invariance, which requires that two versions of the same choice problem that are recognized as equivalent when shown together should elicit the same response when shown separately.[23] For suppose we are asked to evaluate each painful episode E_i as follows:

> "Imagine that you have two possible futures. The first is to live for precisely another T years in good health and without experiencing any significant episode of pain except for the fact that, starting tomorrow, you will have to experience episode E_i.[24] The second is to live for time T_i* in good health and without having to experience E_i or any other significant episode of pain. How long would T_i* have to be to render you indifferent between these two futures?[25]
>
> You may refuse to answer if you feel unable to come up with an answer you regard as even somewhat reliable."

This question is a variant of a method of assessing the value of health states known as the Time Trade-off Method. This method is widely accepted in health economics because, compared to other methods of eliciting people's evaluations of health states, it performs relatively well on the following four central criteria: feasibility (since subjects are generally capable of answering questions of this type), discriminative power (since subjects are able to discriminate health states that differ only slightly), reliability (since subjects tend to give similar answers when asked the question again) and validity (since tests indicate that it accurately reflects the concept it is intended to measure).[26] It is therefore regarded by many as a practical gold standard of health state assessment.[27] It is also a particularly good measure for our purposes, since answers to this question will probably not suffer from the two biases associated with the initial presentation of the decision problem. Unlike similarity-based decision-making, it requires us to engage in a global evaluation of the badness of each episode separately.[28] This global evaluation is carried out in terms of a quantity that we will probably be familiar with and which we are used to making choices about, since our lives regularly involve choices which trade off particular goods against the time we can expect to spend in a pain-free state of full health. Second, because the question requires us to work with duration, this aspect of the episodes will appear salient to us and be rendered more easily evaluable. This will prevent the underweighting of duration associated with

methods of evaluation that give a representative moment from the episode great weight. For these reasons, I will assume that asking the Time Trade-off question is a legitimate way of eliciting a person's preferences in this case. What can we learn from our attempts to answer it?

Suppose first that, perhaps after some practice, we are capable of answering the Time Trade-off question for all alternatives. On this measure, an episode E_i will be better than an episode E_j if and only if T_i* is greater than T_j*. Now, at least one of the pairwise preferences generated by this way of evaluating the same alternatives will therefore be different from the pairwise preferences expressed in the initial presentation of the alternatives. The location of this difference or these differences will be informative. If the only differences are located among adjacent alternatives, then this suggests that similarity-based decision-making led us to undervalue the intensity of pain attribute in our choices between these alternatives. If, by contrast, the Time Trade-off measure agrees with all our initial preferences over adjacent alternatives, and differs only in the assessment of E_0 and E_{MILD}, then this suggests that we underweighted duration in that choice, and that similarity-based decision-making may even have helped us avoid underweighting duration in our other choices. Finally, if the Time Trade-off measure disagrees with some of our initial preferences between adjacent alternatives and with our initial preference between the first and final alternative, then this suggests that our initial decision-making was subject to both aforementioned biases, and that we underweighted the intensity of pain in some choices, and underweighted duration in another choice.

Suppose now, by contrast, that we are unable to answer the Time Trade-off question for some alternatives, because we do not feel that we can accurately globally evaluate some of the alternatives. Since we would now express no preference over some of the alternatives whereas we did express such a preference in the initial presentation of the example, our preferences will again violate invariance. Moreover, our inability to assess some alternatives would suggest that our initial preferences involving these alternatives were induced by the fact that the initial

presentation of the decision-problem enabled us to avoid such global evaluation, instead allowing us to proceed in the comparison of adjacent alternatives by intra-aspect evaluation only (as in similarity-based decision-making), or, in the comparison of the first and last alternatives, by a mental representation of the alternatives that focused our attention inordinately on the intensity of pain involved in each alternative. We should conclude that we do not really know whether these alternatives are better or worse than other alternatives.

Acyclicity's intuitive appeal and its central role in the theory of choice imply that one requires considered, confidently held case judgments that violate acyclicity before one can have sufficient grounds for rejecting it. The preceding considerations establish, I believe, that the intuitive preferences elicited by Rachels' and Temkin's example should not be held with a high degree of confidence. For we know that they may be the result of one or more well-documented biases, and that they will differ from the preferences revealed via a method of preference elicitation that is regarded by experts as among the most reliable of health state evaluation methods. The intuitive preferences elicited by this example should not, therefore, count as grounds for rejecting acyclicity.

Interestingly, psychological research indicates that even if we are persuaded by this argument about the unreliability of our initial preferences in Rachels' and Temkin's example, this may not modify our gut feelings about this example. As Kahneman notes, the operation of intuitive methods of decision-making is typically "fast, automatic, effortless, associative, and difficult to control or modify", and the judgments that these methods generate may therefore be hard to shake off.[29] The pull of our initial preferences in this case may therefore be strong and persistent. It is, nonetheless, a pull we should resist.

---

[1] This paper was presented in the LSE Choice Group Seminar in June 2007. I am grateful to those present, and to Paul Anand, Nick Baigent, Ken Binmore, Richard Bradley, Erik Carlson, Marco Mariotti, Alex Oliver, Michael

Otsuka, Stuart Rachels, Ariel Rubinstein and especially Jean-Francois Bonnefon for comments and/or discussions on the topics addressed in this paper.

[2] See Amarya Sen, *Collective Choice and Social Welfare* (San Fransisco: Holden Day), pp. 15-6.

[3] See Stuart Rachels, *A Theory of Beneficence* (unpublished undergraduate thesis, University of Oxford, 1993); 'Counterexamples to the Transitivity of Better Than', *Australasian Journal of Philosophy* 76 (1998): 71-83; 'A Set of Solutions to Parfit's Problems', *Nôus* 35 (2001): 214-38; and 'Intransitivity', in Volume II of *The Encyclopedia of Ethics* (2nd edition), Lawrence Becker and Charlotte Becker, eds. (London: Routledge, 2001), pp. 877-9; Larry Temkin, 'A Continuum Argument for Intransitivity', *Philosophy & Public Affairs* 25 (1996): 175-210. Rachels' and Temkin's counterexamples are directed at transitivity, which requires that when an alternative a_1 is better than a_2, …, a_{i-1} is better than a_i, …, a_{n-1} is better than a_n, then a_1 is better than a_n, irrespective of the length n of the chain. I have phrased the argument in terms of acyclicity because it is weaker than transitivity and is commonly regarded as the weakest requirement of rational choice.

[4] For similarity-based decision-making, see Amos Tversky, 'Intransitivity of Preferences', *Psychological Review* 84 (1969): 327-52 and Ariel Rubinstein, 'Similarity and Decision-Making Under Risk', *Journal of Economic Theory* 46 (1988): 145-53. Research on the evaluation of painful episodes is reviewed in Daniel Kahneman, 'Maps of Bounded Rationality: A Perspective on Intuitive Judgment and Choice', *Nobel Prize Lectures* (2002): 449-89; Dan Ariely and George Loewenstein, 'When Does Duration Matter in Judgment and Decision Making?', *Journal of Experimental Psychology: General* 129 (2000): 509–23; and Ariely, Kahneman and Loewenstein, 'Joint Comment on 'When Does Duration Matter in Judgment and Decision Making?'', *Journal of Experimental Psychology: General* 129 (2000): 524–9.

[5] Rachels and Temkin regard their argument as resting on three general claims, which they maintain, together entail a violation of acyclicity (see Rachels, 'A Set of Solutions', pp. 215f. and Temkin, 'A Continuum Argument', pp. 179 and 182ff.):

> *Claim 1*: For any unpleasant experience, no matter what the intensity and duration of that experience, it would be better to have that experience than one that was only a little less intense but that lasted much longer.
>
> *Claim 2*: There is a finely distinguishable range of unpleasant experiences ranging in intensity from extreme agony to mild discomfort.
>
> *Claim 3*: No matter how long it must be endured, mild discomfort is preferable to extreme agony for a significant amount of time.

As Ken Binmore and Alex Voorhoeve point out in 'Defending Transitivity Against Zeno's Paradox', *Philosophy & Public Affairs* 31 (2003): 272-9, these three claims do not entail a violation of acyclicity. For it is possible for a person's preferences to be acyclic and satisfy claims 1, 2, and 3. For example, a person who maximizes the utility function

$$u(p,t) = \frac{-pt}{(1+t)}$$

Where u is utility, $p \geq 0$ is the intensity of pain, and $t \geq 0$ is the length of time it must be endured, satisfies the three claims and has acyclic preferences. However, as Erik Carlson has pointed out in 'Intransitivity Without Zeno's Paradox', in *Recent Work in Intrinsic Value*, T. Rønnow-Rasmussen and M. Zimmerman (eds.) Dordrecht: Springer (2005): 273-7, this flaw is easily remedied by replacing Claim 1 with:

> *Claim 1**: Take a sequence of pain experiences ranging from extreme agony to mild discomfort in which each pain in the sequence is slightly less intense than its predecessor. For any pain experience in this sequence, no matter what the intensity and duration of that experience, it would be better to have that experience than the following level of pain in the sequence for a much longer time.

As outlined in the main text, in this revised argument the difficulty lies with Claim 3, because our intuitive judgments over periods of time that far exceed human experience cannot be trusted. Rachels' and Temkin's argument is therefore better regarded as resting on the claim that it is possible to construct a sequence of bundles of pain and duration in which no alternative involves excessively long periods of time and over which our confidently held preferences violate acyclicity. This claim is the target of this paper.

[6] Rachels, 'Solutions', p. 218; Temkin, 'Continuum', pp. 191ff.

[7] 'Continuum', pp. 194-5.

[8] See Alastair Norcross, 'Comparing Harms: Headaches and Human Lives', *Philosophy & Public Affairs* 26 (1997): 135-67. Rachels and Temkin anticipate these worries, and respond that it will be possible to construct a sequence of the requisite kind without appealing to inordinately long periods of time. See Rachels, 'Counterexamples', p. 74 and Temkin, 'Continuum', pp. 184ff.

[9] See Rubinstein, 'Similarity'; and Tversky 'Intransitivity'.

[10] See, Tversky, 'Intransitivity'; Barbara Mellers and Karen Biagini, 'Similarity and Choice', *Psychological Review* 101 (1994): 505-18; Jonathan Leland, 'Generalized Similarity Judgments: An Alternative Explanation for Choice

Anomalies', *Journal of Risk and Uncertainty* 9 (1994): 151-72; Bob Raynard, 'Reversals of Preference Between Compound and Simple Risks: The Role of Editing Heuristics', *Journal of Risk and Uncertainty* 11 (1995): 159-75; David Buschena and David Zilberman, 'Performance of the Similarity Hypothesis Relative to Existing Models of Risky Choice', *Journal of Risk and Uncertainty* 11 (1995): 233-62 and 'Testing the Effects of Similarity on Risky Choice: Implications for Violations of Expected Utility', *Theory and Decision* 46 (1999): 251-76; and Rubinstein, 'Economics and Psychology? The Case of Hyperbolic Discounting', *International Economic Review* 44 (2003): 1207-16. There also appears to be support for the hypothesis that non-human animals use this heuristic. See Sharoni Shafir, 'Intransitivity of Preferences in Honey Bees: Support for 'Comparative' Evaluation of Foraging Options', *Animal Behaviour* 48 (1994): 55-67. Might a honey bee whose preferences violate acyclicity due to their use of similarity-based decision-making be at risk of being turned into a honey pump?

[11] See Amos Tversky and Daniel Kahneman, 'Extensional vs. Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment', *Psychological Review* 90 (1983): 293-315 and 'Prospect Theory', *Econometrica* 47 (1979): 263-91.

[12] See Tversky, 'Intransitivity', p. 42. The general tendency for individuals to avoid inter-dimensional trade-offs is discussed in Eduard Brandstätter, Gerd Gigerenzer, and Ralph Hertwig, 'The Priority Heuristic: Making Choices without Trade-Offs', *Psychological Review* 113 (2006): 409-32. Paola Manzini and Marco Mariotti provide an interesting model of decision-making that avoids trade-offs in 'Sequentially Rationalizable Choice', *American Economic Review*, forthcoming.

[13] See Tversky, 'Intransitivity', p. 43.

[14] See Rubinstein, 'Similarity', and Xavier Vilà, 'On the Intransitivity of Preferences Consistent with Similarity Relations', *Journal of Economic Theory* 79 (1998): 281-7.

[15] Tversky, 'Intransitivity', pp. 37-40.

[16] See Ariely and Loewenstein, 'Duration' and Ariely, Kahneman and Loewenstein, 'Comment'.

[17] Redelmeier and Kahneman, 'Patients' Memories of Painful Medical Treatments: Real-time and Retrospective Evaluations of Two Minimally Invasive Procedures', *Pain* 66 (1996): 3-8.

[18] Support for this hypothesis also appears in Carol Varey and Daniel Kahneman, 'Experiences Extended across Time: Evaluation of Moments and Episodes', *Journal of Behavioral Decision Making*, 5 (1992): 169-85. See also the literature cited in Daniel Kahneman, Peter Wakker, and Rakesh Sarin, 'Back to Bentham? Explorations of Experienced Utility', *Quarterly Journal of Economics* 112 (1997): 375-405.

[19] See Charles Schreiber and Daniel Kahneman, 'Determinants of the Remembered Utility of Aversive Sounds' *Journal of Experimental Psychology: General* 129 (2000): 27-42 and Kahneman, 'Maps of Bounded Rationality', pp. 477-8.

[20] See the literature cited in Kahneman, Wakker and Sarin, 'Back to Bentham?' and Schreiber and Kahneman, 'Determinants'.

[21] See Ariely and Loewenstein, 'Duration', p. 513.

[22] For an argument to this effect in a different thought-experiment, see Alex Voorhoeve and Ken Binmore, 'Transitivity, Similarity-Based Decision-Making, and the Sorites Paradox', *Erkenntnis* 64 (2006): 101-14.

[23] For a discussion of this principle, see, for example, Kahneman and Tversky, 'Choices, Values and Frames', *American Psychologist* 39 (1984): 341-50.

[24] Period T must be at least as long as the duration of the pain in E_{MILD}.

[25] As we assumed at the outset, though enduring episode E_i is bad, living for period T and having to endure episode E_i during some part of that period T is better than dying immediately, so that T_i^* will be larger than 0.

[26] For discussion of the Time Trade-off Method, see Paul Dolan, 'Output Measures and Valuation in Health', in Michael Drummond and Alistair McGuire (eds.) *Economic Evaluation in Health Care* (Oxford: Oxford University Press, 2001), pp. 46-67 and Sylvie van Osch, Peter Wakker, Wilbert van den Hout and Anne Stiggelbout, 'Correcting Biases in Standard Gamble and Time Tradeoff Utilities', *Medical Decision Making* 24 (2004): 511-7.

[27] Van Osch *et. al.* 'Correcting Biases', p. 515.

[28] To avoid our answers being influenced by our answers to episodes involving similar intensities of pain, we should ensure that each episode that is evaluated is significantly different along both dimensions from the episode that preceded it in evaluation. This would involve, for example, first evaluating E_4, then E_9, then E_2, etc.

[29] 'Maps of Bounded Rationality', p. 450.